Dialogue and Discourse 3(2) (2012) 177–204 doi: 0.5087/dad.2012.208

A Detailed Account of The First Question Generation Shared Task Evaluation Challenge

Vasile Rus VRUS@MEMPHIS.EDU
Department of Computer Science
The University of Memphis
USA

Brendan Wyse BJWYSE@GMAIL.COM
AOL, Dublin
Ireland

Paul Piwek P.PIWEK@OPEN.AC.UK
Centre for Research in Computing
Open University, UK

Mihai Lintean MCLINTEN@MEMPHIS.EDU
Department of Computer Science
The University of Memphis
Memphis, TN 38152
USA

Svetlana Stoyanchev S.STOYANCHEV@OPEN.AC.UK
Centre for Research in Computing
Open University, UK

Cristian Moldovan CMLDOVAN@MEMPHIS.EDU
Department of Computer Science
The University of Memphis
Memphis, TN 38152
USA

Editor: Kristy Elizabeth Boyer

Abstract

The paper provides a detailed account of the First Shared Task Evaluation Challenge on Question Generation that took place in 2010. The campaign included two tasks that take text as input and produce text, i.e. questions, as output: Task A – Question Generation from Paragraphs and Task B – Question Generation from Sentences. Motivation, data sets, evaluation criteria, guidelines for judges, and results are presented for the two tasks. Lessons learned and advice for future Question Generation Shared Task Evaluation Challenges (QG-STEC) are also offered.

Keywords: question generation, shared task evaluation campaign.

 Submitted 3/11; Accepted 1/12; Published online 3/12

1 Introduction

Question Generation is an essential component of learning environments, help systems, information seeking systems, multi-modal conversations between virtual agents, and a myriad of other applications (Lauer, Peacock, and Graesser, 1992; Piwek et al, 2007).

Question Generation has been recently defined as the task of automatically generating questions from some form of input (Rus & Graesser, 2009). The input could vary from information in a database to a deep semantic representation to raw text. Question Generation is viewed as a three-step process (Rus & Graesser, 2009): content selection, selection of question type (Who, Why, Yes/No, etc.), and question construction. Content selection is about deciding what the question should be about given the various inputs available to the system and the context in which the question is being asked. In other words, this step focuses on deciding what is important to ask about in a particular context. The step of question type selection decides the most appropriate type of the question given the selected content and context. That is, this step focuses on how to ask the question. For instance, a *how* or *why* question type may be selected or even a *yes/no* type of question. The last step, question construction, is about realizing the question given the selected content, question type, and the context in which the question is being asked. This step too focuses on how to ask the question. A fourth step has been proposed by Mostow and Chen (2009) which refers to the decision of when to ask a question in a particular context. A question has its most desired effect when asked at the right moment. While finding the right moment seems to be an open research challenge at this moment, preliminary work has noted that a question should not be asked "too soon after the prior one" in the context of reading comprehension assessment or scaffolding (Beck, Mostow, & Bey, 2004; Mostow et al., 2004).

Question Generation (QG) is primarily a dialogue and discourse task. It draws on and overlaps with research in both Natural Language Understanding (NLU) and Natural Language Generation (NLG). For example, QG from raw text, e.g. a textbook or the dialogue history between a tutor and tutee, will involve some sort of analysis, i.e., NLU, of the input to select content, detect patterns, etc. The construction of the output question will typically involve NLG: a representation of the input is mapped to a sentence. QG from non-linguistics input knowledge representations (e.g., Semantic Web ontology languages), can be viewed directly as a typical NLG task that takes as input non-linguistic data and produces output in a human language (Reiter & Dale, 1997; Reiter & Dale, 2000).

QG as a focused area of research is relatively young. It basically started in 2008 with the National Science Foundation-sponsored workshop on The Question Generation Shared Task and Evaluation Challenge[1] (Rus & Graesser, 2009). Sporadic research on Question Generation did happen before (Wolfe, 1976; Kunichika et al., 2001; Mitkov, Ha, & Karamanis, 2006; Rus, Cai, & Graesser, 2007a). The QG research community has decided to offer shared task evaluation campaigns or challenges (STECs) as a way to stimulate research and bring the community closer.

The first Question Generation Shared Task Evaluation Challenge (QG-STEC) follows a long tradition of STECs in Natural Language Processing: see various tracks at the Text REtrieval Conference[2] (TREC), e.g. the Question Answering track (Voorhees & Tice, 2000), the semantic evaluation challenges under the SENSEVAL[3] umbrella (Edmonds, 2002), or the annual tasks run by the Conference on Natural Language Learning[4] (CoNLL). In particular, the idea of a QG-STEC was inspired by the recent activity in the Natural Language Generation (NLG) community to offer shared task evaluation campaigns as a potential avenue to provide a focus for research in

[1] www.questiongeneration.org

[2] http://trec.nist.gov

[3] www.senseval.org

[4] http://www.cnts.ua.ac.be/conll/

NLG and to increase the visibility of NLG in the wider Natural Language Processing (NLP) community (Dale and White, 2007). The NLG community has offered under the umbrella of Generation Challenges[5] several shared tasks on language generation including generation of natural-language instructions to aid human task-solving in a virtual environment (the GIVE-2 challenge[6]; Koller et al., 2010) and post-processing referring expressions in extractive summaries (GREC'10; Belz et al., 2008; Belz & Kow, 2010).

In designing the first QG-STEC, we had to balance conceptual and pragmatic issues. Two core aspects of a question are the goal of the question and its importance. It is difficult to determine whether a particular question is good without knowing the context in which it is posed; ideally one would like to have information about what counts as important and what the goals are in the current context. This suggests that a STEC on QG should be tied to a particular application, e.g. tutoring systems. However, an application-specific STEC would limit the pool of potential participants to those interested in the target application. Therefore, the challenge was to find a framework in which the goal and importance are intrinsic to the source of questions and less tied to a particular context/application. One possibility was to have the general goal of asking questions about salient items in a source of information, e.g. core ideas in a paragraph of text. Our Task A (described later) has been defined with this concept in mind. This basic idea does not really hold for generation of questions from single sentences, which is our Task B (defined later). For Task B, we have basically decided to ignore the importance of the question and chose to accept all questions as long as they fit some other minimal criteria for a good question (fluency, ambiguity, relevance), which are application independent. Ignoring the importance of the questions for Task B seemed appropriate given that the research on this type of systems is still in a very early phase. Task B focused more on evaluating the capacity of systems to construct questions rather than select content for asking questions about.

Adopting the basic principle of application-independence has the advantage of escaping the problem of a limited pool of participants (to those interested in a particular application had that application been chosen as the target for a QG-STEC).

Besides the advantage of a larger pool of potential participants, an application-independent QG-STEC would provide a more fair ground for comparison as teams already working on a certain application would not be advantaged as would be the case if the application had been the focus of a STEC. It should be noted that the idea of an application-independent STEC is not new. An example of an application-independent STEC would be generic summaries (as opposed to query-specific summaries) in summarization.

Another decision aimed at attracting as many participants as possible and promoting a more fair comparison environment concerned the input for the QG tasks. A particular semantic representation would have provided an advantage to groups already working with it and at the same time it would have raised the barrier-to-entry for newcomers. Instead, we have adopted a second guiding principle for the first QG-STEC tasks: no representational commitment. That is, we wanted to have as generic an input as possible. Therefore, the input to both task A and B in the first QG-STEC was raw text. That is, the first QG-STEC falls in the wider category of Text-to-Questions tasks identified at the First Workshop on Question Generation (Rus & Graesser, 2009). Also, it can be viewed as a Text-to-Text tasks as proposed by Rus and colleagues (2007b).

Task A and B offered in the first QG-STEC fall in the Text-to-Question category of QG tasks identified by The First Workshop on Question Generation (www.questiongeneration.org). The first workshop identified four categories of QG tasks (Rus & Graesser, 2009): Text-to-Question, Tutorial Dialogue, Assessment, and Query-to-Question. Using another categorization, tasks A

[5] http://www.itri.brighton.ac.uk/research/genchal10/

[6] http://www.give-challenge.org/

and B are part of the Text-to-text Natural Language Generation task category identified by the Natural Language Generation community (Rus et al., 2007b).

There was overlap between Tasks A and B in the first QG-STEC. This was intentional with the aim of encouraging people preferring one task to also participate in the other. In particular, generating questions from individual sentences (Task B) was also included as one goal among several others in Task A (where the input consisted of paragraphs).

We opted for human-based evaluation for the first QG-STEC. Evaluation in Question Generation is closest to evaluations in natural language generation, machine translation, and summarization, as the outputs in all these areas correspond to texts in a human language, similar to the Question Generated shared tasks in the first QG-STEC. In machine translation and summarization, the input is also text, as for the tasks in the first QG-STEC. While in machine translation and summarization automatic scoring procedures are nowadays acceptable as they have been shown to correlate with human ratings (Papineni et al., 2002; Lin, 2004; Lin and Och, 2004), in natural language generation human-based evaluations are the current norm (Walker, Owen, & Rogati, 2002). For the time being, we have opted for the conservative option of relying on human judgment for both conceptual and pragmatic reasons. As in natural language generation, for a given input text there is in theory a large number of questions that can be asked about. In other words, there is no uniquely and clearly defined correct answer for Question Generation. Additionally, given the early stage of research in Question Generation, the safe approach to evaluation would be the human-based approach. Automatic scoring is an interesting topic of future research for which the first QG-STEC can be used as a testbed. For certain types of questions, automatic scoring is possible based on extrinsic criteria. For instance, the quality of the automatically generated multiple-choice questions by Mitkov, Ha, and Karamanis (2006) were evaluated using item test theory. This scoring resembles the task-based (extrinsic) form of evaluation used in natural language generations in some cases. Furthermore, at the time of planning the first QG-STEC there was only one previous work, to the best of our knowledge, that evaluated questions of the type we focused on, i.e. non multiple choice questions, generated from raw text (Rus, Cai, & Graesser, 2007a). In line with most existing evaluation schemes, for the QG-STEC the focus was on measuring the quality of the output (i.e., the generated questions) along several dimensions. Such an absolute measure does, however, not take into account the quality of the input. Ideally, especially for criteria such as fluency, one may want to measure to what extent the score for the output question is an increase or decrease relative to the score for the input text, e.g., is the fluency of the output question better or worse than that of the input text and how much better or worse? Such an approach measuring relative quality of outputs has been proposed and applied by Piwek & Stoyanchev (2011) to the task of generating short fragments of dialogue (e.g., question-answer pairs) from expository monologue.

Overall, we had one submission for Task A and four submissions for Task B. The submissions were evaluated through a peer-review system for Task B. Task A was evaluated by two external judges as there was only one submission and the peer-review mechanism could not be applied. There are several explanations for the relatively small number of participants. First, there are no well-established research programs in this area which would allow researchers to dedicate their time to develop competitive QG systems and participate in QG-STECs. This argument is further supported by the discrepancy between the number of research groups who expressed interest in participating and those who really submitted results for evaluation. In particular for Task A, the discrepancy was quite significant: five teams expressed explicit interest to participate (as required by the QG-STEC organizers, but only one team submitted results for evaluation. In a way, Task A is more difficult than Task B (it requires discourse level processing), which explains the lower number of submitted teams, although the number of interested teams was comparable. Second, being the first QG-STEC ever offered, the effort necessary to competitively participate was substantial. In subsequent QG-STECs, participating teams will be able to build on the tools and insight generated either by themselves or other teams that

participated in the first QG-STEC. Third, in future QG-STECs publicity for the STEC could be improved, perhaps by aligning with the Generation Challenges initiative or another mainstream shared task initiative, and by allowing for more time between release of the STEC instructions/development data and the release of test data. All the data together with detailed descriptions of the two tasks and the papers describing the approaches proposed by the participants are accessible from the main question generation website (see footnote 1).

It is important to say that the two tasks offered in the first QG-STEC were selected among five candidate tasks by the members of the QG community. A preference poll was conducted and the most preferred tasks, Question Generation from Paragraphs (Task A) and Question Generation from Sentences (Task B), were chosen to be offered in the first QG-STEC. The other three candidate tasks were: Ranking Automatically Generated Questions (Michael Heilman and Noah Smith), Concept Identification and Ordering (Rodney Nielsen and Lee Becker), and Question Type Identification (Vasile Rus and Arthur Graesser). The involvement of the QG community at large contributed to quality of tasks offered and ultimately the success of the first QG-STEC.

2 Task A: Question Generation from Paragraphs

In this section, we present the details of Task A, Question Generation from Paragraphs, together with the results of the participating system. All the data and guidelines are accessible from the Question Generation wiki[7].

2.1 Task Definition

The Question Generation from Paragraphs (QGP) task challenged participants to generate a list of questions from a given input paragraph. The questions should be at three specificity/scope levels: general/broad (triggered by the entire input paragraph), medium (one or more clauses or sentences), and specific (phrase level or less). Participants were asked to generate one general question, two medium questions, and three specific questions for a total of six questions per input paragraph. The specificity/scope was defined by the portion of the paragraph that answered the question. If multiple questions could be generated at one level, only the specified number should be submitted. For example, for a paragraph that answers two broad questions, only one question at that level of specificity should be submitted.

For the QGP task, questions are considered important and interesting if they ask about the core idea(s) in the paragraph and an average person reading the paragraph would consider them so, based on a quick analysis of the contents of the paragraph.

Simple, trivial questions such as *What is X?* or generic questions such as *What is the paragraph about?* were to be avoided. In addition, implied questions (whose answer is not explicitly stated in the paragraph) were not allowed as the emphasis was on questions triggered and answered by the paragraph. Furthermore, questions could not be compounded as in *What is ... and who ... ?* Questions had to be grammatically and semantically correct and related to the topic of the given input paragraph. Question types (*who/what/why/...*) generated for each paragraph should be diverse, i.e. different question types are preferred in the set of 6 required questions.

2.2 Guidelines for Human Judges

We next show an example paragraph together with six interesting, application-independent questions that could be generated. The paragraph is about two-handed backhands in tennis and was collected from Wikipedia. Table 1 shows the paragraph while Table 2 shows six questions triggered by this paragraph. Each question is at a different level of specificity. The first question

[7] http://www.questiongeneration.org/mediawiki/index.php/QG-STEC_2010

is general/broad because its answer is the entire paragraph. One may argue that the first sentence in the paragraph, which usually is the topic sentence summarizing the paragraph, also forms a valid answer to the question. This is true in general. For this reason, we instructed the judges to consider the widest scope possible when judging a question and also to judge the question within the scope indicated by participants. In other words, if the participants selected the entire paragraph as opposed to just the topic sentence as the content that triggered the question then judges should rate the question scope within the indicated scope: Does the question fit the participant-selected scope?

Table 1. Example of input paragraph (from http://en.wikipedia.org/wiki/Backhand) together with the answers to the questions shown in Table 2.

Input Paragraph
Two-handed backhands have some important advantages over one-handed backhands. Two-handed backhands are generally more accurate because by having two hands on the racquet, this makes it easier to inflict ***topspin*** *on the ball* **allowing for more control of the shot**. *Two-handed backhands are easier to hit for most* ***high balls***. *Two-handed backhands can be hit with* ***an open stance***, *whereas one-handers usually have to have a closed stance, which adds further steps (which is a problem at higher levels of play).*

Table 2. Examples of questions and scores for the paragraph in Table 1. The fragment in between [] is optional.

Questions	Scope
What are the advantages of two-handed backhands in tennis?	*General*
Why are two-handed backhands more accurate [when compared to one-handers]?	*Medium*
What is one consequence of inflicting topspin on a tennis ball?	*Medium*
What kind of spin does a two-handed backhand inflict on the ball?	*Specific*
What stance is needed to hit a two-handed backhand?	*Specific*
What types of balls are easier to hit with a two-handed backhand?	*Specific*

The next two questions in Table 2 are medium-scope questions as their answers are entire sentences (see the underlined and bold sentences, respectively, in the input paragraph shown in Table 1 which form the answers to the two questions).

The bottom three questions in Table 2 are specific questions whose answers are a phrase or less in length. The phrase answers to the example specific questions are shown in bold face in the input paragraph in Table 1.

The three specificity levels proposed for Task A were consciously chosen by the proposing team as explained next. The specific questions were inspired by the factoid questions used for shared tasks by the Question Answering community (Voorhees & Tice, 2000). The answer to these factoid questions were usually short snippets of text in the form of a phrase or even less, e.g. the name of a person such as President Obama. Examples of factoids questions used by the Question Answering STECs are *Who is the voice of Miss Piggy?*, whose answer is "Frank Oz", and *How much could you rent a Volkswagen bug for in 1966?* whose answer is "$1/day". By focusing on specific facts in the input paragraph, QG-STEC participants could be challenged to generate specific questions. Additionally, these questions corresponded closely to the questions

required for Task B (on QG from sentences). The medium-scope questions were inspired by discourse-level relations which usually hold among larger chunks of texts in the input paragraph (see Piwek et al., 2007 for automatic generation of such questions in expository dialogue). As an example, we use the cause-effect relation between the underlined versus the simple italic text portion of the sentence "Two-handed backhands are generally more accurate because by having two hands on the racquet, this makes it easier to inflict topspin on the ball allowing for more control of the shot." (from the input paragraph in Table 1). The discourse relations could be used to generate questions where one fragment in the discourse relation forms the body of the question while the other fragment forms the answer to the question, i.e. the target content triggering the question. The cause-effect relation example provided above can be used to trigger a *Why* question whose body is the effect portion of the relation (*Why are two-handed backhands more accurate*?) while the answer is the cause portion of the relation: "because by having two hands on the racquet, this makes it easier to inflict topspin on the ball allowing for more control of the shot." The general/broad scope questions could be generated by doing a global analysis of the information content of the paragraph based on which the more salient concepts could be used to trigger such general/broad questions whose answer is the entire paragraph.

We now turn to describing the evaluation criteria used to judge submitted questions. These criteria were used by human raters to judge the questions. We will use the paragraph and questions in Tables 1 and 2 to illustrate the judging criteria.

A set of five scores, one for each of the following criteria was generated for each question.

- Specificity
- Syntax
- Semantics
- Question type correctness
- Diversity

Additionally, we anticipated creating a composite score based on the scores for the individual criteria. This score would range from 1 (first/top ranked, best) to 4 (lowest rank), with 1 meaning that the best possible score was achieved on each of the individual criteria. However, deciding on a valid combination of the individual scores is extremely difficult (e.g., are the criteria all of the same weight?), whilst it would force a single ranking on the performance of participating systems. For these reasons, we decided not to calculate such a composite score, but keep this on the agenda as an issue for further discussion in the run-up to the next QG-STEC for Task A. The specificity scores are assigned primarily based on the answer span in the input paragraph. The broadest question is the one whose answer spans the entire paragraph. The most specific question is the one whose answer is less than a sentence: a clause, phrase, word, or collocation. Scores were assigned based on the following rubric: 1 – input paragraph, 2 – multiple sentences, 3 – a clause or less, 4 – trivial/generic, implied, no question (empty question), or undecided, e.g. a semantically wrong question which may not be understood well enough to judge its scope. As we expected six questions as output, if one level is missed we encouraged participants to generate questions of a narrower scope. For instance, if a broad-scope question could not be generated then a medium or specific question should have been submitted. This assured that participants submitted as many as six questions for each input paragraph.

The best question specificity scores for six questions would be 1, 2, 2, 3, 3, 3. This best configuration of scores would be possible for paragraphs that could trigger the required number of questions at each scope level, which may not always be the case. In preparing the data for the QG-STEC, we aimed at selecting paragraphs that made it possible to obtain perfect scope scores.

While the initial plan was for the judges to look at the question itself and select themselves the portion of the paragraph that may have triggered the question, we opted instead, for practical reasons, to allow the judges to see the span of text submitted by participants for each question and decide based on the participant-submitted span the specificity of the question. The advantage of

the initial plan is that the judges' selected text span could have been automatically compared to the span submitted by participants for an automated scoring process of the scope criterion. However, this initial plan proved to be more challenging because if judges were allowed to select their own answer span without seeing the one provided by participants then that may have undesired effects on human-judging of the questions on the other criteria, e.g. question type correctness or semantic correctness. For instance, it may be the case that a judge selects a slightly different span for the question than the one targeted by the participant, which may imply a question type different from the one chosen by the participants. Another problem occurs for questions that have semantic issues. For such questions, guessing the fragment in the input paragraph that triggered them can be challenging for the human judges. Having the fragment already available as indicated by participants would allow the judges to better evaluate the question along all criteria.

The semantic correctness was judged using the following scores: 1 – semantically correct and idiomatic/natural, 2 – semantically correct and close to the text or other questions, 3 – some semantic issues, 4 – semantically unacceptable. Table 3 shows examples of real questions submitted by The University of Pennsylvania team and the corresponding semantic correctness scores. Scores of 3 and 4 were relatively easy to assign. Sometimes, it was harder to differentiate between scores of 1 and 2 because while the question seems natural and almost impossible to formulate in a more natural way, it was too close or identical to the text.

Table 3. Examples of questions corresponding to different levels of semantic correctness.

Questions	Semantic Score
What is the porosity of an aquifer?	*1*
How are diamonds brought close to the earth surface by a magma, which cools into igneous rocks known as kimberlites and lamproites?	*2*
Who is ibn sahl credited with?	*3*
What might she to spend like his other childless concubines as?	*4*

The syntactic correctness was judged using the following scores: 1 – grammatically correct and idiomatic/natural, 2 – grammatically correct, 3 – some grammar problems, 4 – grammatically unacceptable. Table 4 shows examples of real questions submitted by participants and the corresponding syntactic correctness scores. Scores of 3 and 4 were easy to assign. Differentiating between scores of 1 and 2 was at times difficult because the naturalness of the question is more difficult to judge when there is no obvious more natural way to ask a question given a particular target fragment in the input paragraph.

Correctness of question type means the specified type by a participant is agreed upon by the human judge. This is a binary dimension: 1 – means the judge agrees with the specified answer type, 0 – means the judge disagrees. An example of a wrong question type is provided in the following
question, *Who do we see in the next scene?*, whose correct type was *What*. Question type correctness depends on the target content that triggered the question, as well as the body of the question. Again, we assumed the target content submitted by the participants as being fixed/correct and judged the question type correctness with this assumption in mind. There are several cases that had to be considered. First, the question type was deemed incorrect when it did not match the target content and the question body. Second, when the question body is semantically unacceptable, the question type can either be considered incorrect by default or can be judged correct if the target content does imply the selected question type. We chose to judge the correctness of question type with respect to the target content, ignoring the question body

when semantically unacceptable. The reason for this choice was to avoid penalizing participants whose question body construction module was less developed. As a final remark on body when

Table 4. Examples of questions corresponding to different levels of syntactic correctness.

Questions	Syntactic Score
What was the immediate cause of the first crusade?	*1*
How are diamonds brought close to the earth surface by a magma, which cools into igneous rocks known as kimberlites and lamproites ?	*2*
What do different materials have a different albedo so reflect a different amount of solar energy?	*3*
What does seek we to apply to life the understanding that separate parts of the ecosystem function as a whole?	*4*

semantically unacceptable. The reason for this choice was to avoid penalizing participants whose question body construction module was less developed. As a final remark on question type correctness, we would like to note that question type is a binary judgment (0-1) as opposed to the other dimensions where we used finer grain ratings (1-4), allowing for some gray ratings instead of just crisp distinctions. This difference in granularity of ratings may explain the differences in score values compared to the scores on the other dimensions.

Diversity of question types was also evaluated. At each scope level, ideally, each question had a different question type. A question type is loosely defined as being formed by the question word (e.g., *wh*-word or auxiliary) and by the head of the immediately following phrase. For instance, in *What U.S. researcher* …? the head of the phrase *U.S. researcher* that follows the question word *What* indicates a person, which means the question is actually a *Who* question and not a *What* question. However, preference was given to diversity of question words. Full question types, i.e. including the head of the phrase following the question word, were considered in special cases when the use of diverse question words was constrained by the input paragraph; that is, when different question words are hard to employ in order to generate different question types. For instance, some paragraphs may facilitate the generation of true *What* questions, i.e. *What* question types, but not *When* questions. For diversity ratings, we will use the following rubric: 1 – diverse in terms of question type and main body, 2 – diverse in terms of main body, 3 – paraphrase of a previous question, and 4 – similar-to-identical to a previous question.

While full diversity would be ideal, it can be quite challenging for some input paragraphs. For pragmatic reasons, we relaxed the diversity criteria. We assigned the highest score of diversity if at least 50% of the question types in the whole 6-question set were different and the distribution of types was balanced, e.g. 2-*Who*, 2-*What*, and 2-*Where* would be scored higher than 4-*Who*, 1-*What*, and 1-*Where*.

We also defined overall average scores for each criterion. The overall average scores were defined as the average of individual scores. An individual score summarizes the scores along a dimension, e.g. syntactic correctness, and is computed by taking the average of individual scores shown by the formula below where Q is the number of questions.

$$Syntactic-overall-score = \frac{\sum_{|Q|} individual\_score}{|Q|}$$

The overall average score for specificity is more challenging to define. The goal would be to have a summative score with values from 1 to 4. 1 should be assigned to perfect system that generates 6 questions for each paragraph with the required distribution of specificity levels: 1 general, 2 medium, and 3 specific. For instance, if there are 4 specific questions in a set of 6 questions, then when judging the fourth specific question it will be penalized because the number

of expected specific questions (3) have been exhausted and another question at general or medium scope has not been generated. As of this writing, we are refining our overall average score for specificity.

2.3 Data Sources and Annotation

The primary source of input paragraphs used for the first QG-STEC were: Wikipedia, OpenLearn,[8] and Yahoo! Answers. The three sources have their own peculiarities. Wikipedia and OpenLearn are collections of well-written texts with Wikipedia being developed in a more ad-hoc manner by volunteer contributors while OpenLearn, a repository of learning materials from Open University courses that has been released for free access by the general public, has been created by professionals, i.e. academics and editors at Open University. Yahoo! Answers is a community-based Question Answering online service in which web surfers ask questions which are then answered by other surfers. Yahoo! Answers contains texts that are less edited resulting in texts that quite often contain ungrammatical sentences. The difference in quality between Wikipedia and Yahoo! Answers texts can be explained by the differences in the way users contribute text and this text is subsequently dealt with. In Wikipedia, once a text is drafted it can be polished by others over iterations until a stable version is reached. In Yahoo! Answers, a first contribution, i.e. answer to a question, is almost never re-edited by another contributor. We chose to use Yahoo! Answers as a source due to the large pool of question-answer pairs available for almost any type of questions (Rus et al., 2009). Eventually, the question-answer pairs can be used to train a system to generate the question given the answer.

We collected (almost) paragraphs from each of these three sources. The paragraphs cover randomly selected topics of general interest. Half of the paragraphs from each source were allocated to a development data set (65 paragraphs) and a test data set (60 paragraphs), respectively. For the development data set, we manually generated and scored 6 questions per paragraph for a total of 6 x 65 = 390 questions.

Paragraphs were selected such that they were self-contained (no need for previous context to be interpreted, e.g. will have no unresolved pronouns) and have around 5-7 sentences for a total of 100-200 tokens (excluding punctuation). In addition, we aimed for paragraphs that can facilitate the generation of 1 broad question, 2 medium questions, and 3 specific questions. Examples of paragraphs from each source are given in Table 5 below.

We decided to provide minimal annotation for input in order to allow individual participants to choose their own preprocessing tools. We did not offer annotations for lemmas, POS tags, syntactic information, or PropBank-style predicate-argument structures. This linguistic information can be obtained with acceptable levels of accuracy from open source tools. This approach favors comparison of full systems in a black-box manner as opposed to more specific components. We did provide discourse relations based on HILDA, a freely available automatic discourse parser (duVerle & Prendinger, 2009).

2.4 Submission Format

For each input paragraph, participants were asked to submit six lines of output. Each line had to contain four items that are tab separated as shown below:

RANK INDEX-SET QUESTION-TYPE QUESTION

where RANK is the rank or identifier of the question starting with 1, INDEX-SET is a set of indices in the input paragraph that indicate the target content in terms of span of tokens (words

[8] OpenLearn gives free access to learning materials from The Open University (http://open.ac.uk/openlearn)

and punctuation) that triggered the question (or can form the answer to the question), QUESTION-TYPE is the type of question, and QUESTION is the generated question. The INDEX-SET should contain pairs of start-end token indices delimited by commas: <0-10, 20-35>. The INDEX-SET itself is delimited by '<' and '>'. QUESTION-TYPE can be one of the following (who, where, when, which, how, how many/long, generic, yes/no; see the Question-from-sentences task description for details regarding these question types) or a new one submitted by participants. Examples of lines in the above format are shown below.

1 <0-145> Who Who is Abraham Lincoln?
2 <98-126> What What major measures did President Lincoln introduce?
3 <66-66> When When was Abraham Lincoln elected president?

Table 5. Examples of input paragraphs from the three sources: Wikipedia, OpenLearn, and Yahoo! Answers.

Source	Paragraph
Wikipedia	*Enzymes are mainly proteins, that catalyze (i.e., increase the rates of) chemical reactions. An important function of enzymes is in the digestive systems of animals. Enzymes such as amylases and proteases break down large molecules (starch or proteins, respectively) into smaller ones, so they can be absorbed by the intestines. Starch molecules, for example, are too large to be absorbed from the intestine, but enzymes hydrolyse the starch chains into smaller molecules such as maltose and eventually glucose, which can then be absorbed. Different enzymes digest different food substances. In ruminants which have herbivorous diets, microorganisms in the gut produce another enzyme, cellulase to break down the cellulose cell walls of plant fiber.*
OpenLearn	*There are two distinct zones containing water beneath the ground surface. The unsaturated zone has mainly air-filled pores, with water held by surface tension in a film around the soil or rock particles. Water moves downwards by gravity through this zone, into the saturated zone beneath, in which all the pores are filled with water. The boundary surface between the unsaturated zone and the saturated zone is the water table, which is the level of water in a well (strictly, in a well that just penetrates to the water table). Water below the water table, in the saturated zone, is groundwater. Just above the water table is a zone called the capillary fringe, in which water has not yet reached the water table, because it has been held up by capillary retention. In this process water tends to cling to the walls of narrow openings. The width of the capillary fringe depends on the size of the pore spaces and the number of interconnected pores. It is generally greater for small pore spaces than for larger ones.*
Yahoo! Answers	*Hi. Depending on the type of duck, some can be trained. Keep them in a dog pen if they are indoors, but with lots of "outside" time. Keeping them out of food and water means getting special dishes just for this. Any good farm supply store will have special dishes which have holes for the beak, and a sloped top so that ducks and poultry can't get into the dish. I'm including a link for the actual "Care" of the duckling. Ducks get big and male ducks, at maturity, can get a little bit mean. If you want a pet duck try to get a female. Make sure to feed appropriate food. Most ducks eat poultry food. Muscovites eat game bird food. Babies need crumbles, and adults get pellets. You can supplement food with scratch grain (but not too much, and not until they are a couple weeks old). Bread should be kept as a once in a while thing. Although, you may find that young ducks have a love of "milk sop". Milk sop is when you soak stale bread in milk so that it's mushy. It's got calcium and protein and while it shouldn't be fed regularly it makes a great treat.*

2.5 Results and Discussion

For Task A, there was one submission out of five registered participants. The participating team was from University of Pennsylvania (Mannem, Prasad, & Joshi, 2010). This section presents their approach and results.

The approach proposed by The University of Pennsylvania team (Mannem, Prasad, and Joshi, 2010) for the task of generating questions from paragraphs (Task A) is an over-generation approach in which many questions are first generated from a myriad of potential content items in the input paragraph followed by a ranking phase. The importance of a question is determined in

two separate steps. First, they use predicate argument structures along with semantic roles to identify important aspects of paragraphs. Required and optional arguments of predicates are considered as good content candidates for asking questions about. Predicates with less than two arguments are excluded and copular verbs are treated differently because of the limitations of the semantic role labeler. It should be noted that the type of arguments is also used for question type selection. This step has limitations when it comes to generating questions that should rely on cross-sentence information, e.g. a cause-effect discourse relation between two sentences. Such cross-sentence information is important for medium and general questions in Task A, as mentioned earlier. Second, in the ranking phase Mannem, Prasad, and Joshi prefer questions generated from main clauses and questions that do not include pronouns, because their approach did not include any coreference resolution. The question formulation step uses the selected content and corresponding verb complex in a sentence together with a set of reformulation rules to generate the actual questions.

The initial plan in terms of evaluating the submissions for Task A was to do peer-reviewing. However, because we only received one submission peer-reviewing was not possible. Instead, we adopted an independent-judges approach in which two independents human raters judged the submitted questions using the interface depicted in Figure 3.

Table 6. Summary of Results for University of Pennsylvania.

Score	Results/Inter-rater Agreement
Specificity	General= 90%; Medium=121%; Specific=80%; Other = 1.39%/68.76%
Syntactic Correctness	1.82/87.64%
Semantic Correctness	1.97/78.73%
Question Diversity	1.85/100%
Question Type Correctness	83.62%/78.22%

Table 7. Summary of Results by Source of Input Paragraphs for University of Pennsylvania.

Score	Wikipedia	OpenLearn	Yahoo! Answers
Specificity	19/46/50/2	19/45/47/1	16/54/47/2
Syntactic Correctness	1.63	1.86	1.98
Semantic Correctness	1.90	1.96	2.08
Question Diversity	1.55	2	1.95
Question Type Correctness	91.73%	87.6%	71.53%

For the 60 input paragraphs used for testing, participants were supposed to submit 60 general questions, 120 medium questions, and 180 specific questions, for a total of 360 questions. The University of Pennsylvania team submitted 349 questions of which 54 were rated general based on the span of the input paragraph fragment that triggered the question (INDEX-SET field in the submission format), 145 as medium, 144 as specific, and 5 as other. The table indicates the percentages of each type of question that were submitted out of what they were supposed to submit. In the *Other* category, we report the percentage of questions that could not be classified as either general, medium, or specific, out of the total number of questions that were supposed to be submitted, i.e. 360. Examples of questions that were classified in the *Other* category are those in which participants submitted an empty span for the input fragment based on which the question was generated, e.g. the INDEX-SET field has a value of 285-285, or the question was so hard to understand that it was impossible to evaluate its scope. The inter-rater agreement, i.e. the percentage of annotated items on which judges agree, on the specificity criterion was 68.76%. The results indicate that participants tended to overgenerate medium questions (they submitted

more questions than asked for) and undergenerate general and specific questions (participants submitted less questions than required).

In terms of syntax, the majority of the submitted questions were deemed correct (score of 1.97 ~ a rate of 2 means grammatically correct) but not necessarily idiomatic or natural. A good majority of the questions were generated by re-using entire chunks of text from the input paragraph, i.e. a selection-based generation approach for the question construction phase.

Semantics-wise, the submitted questions were deemed semantically correct and close to the text and/or other questions (average score of 1.97 ~ 2) but not idiomatic or natural.

The diversity of the submitted questions was rated at 1.85 level, which is close to a score of 2. A score of 2 indicates diversity in terms of the main body of the questions. The types of question and their distribution is shown in the table 8 and charts in Figures 8 and 9 below. It should be noted that the types were obtained by only considering the question word, e.g. *wh*-word, and without any finer-grain analysis, e.g. distinguishing definition questions such as *What is …?* from other *What* questions. We would like to point out that all *What* questions that were submitted are true *What* questions, as opposed to disguised other types as in *What researchers discovered DNA?* which is semantically a *Who* question (asking for a person rather than a thing). That is, all the submitted *What* questions followed the pattern *What auxiliary-verb …?* The Question Type Correctness was 83.62% overall with inter-rater agreement of 78.22%.

Table 8. Distribution of Question types (overall and by input paragraph source).

Question Type	Overall	Wikipedia	OpenLearn	Yahoo! Answers
How	40	16	12	12
What	258	80	87	91
When	22	5	5	12
Where	12	7	4	1
Who	9	5	1	3
Why	8	4	3	1

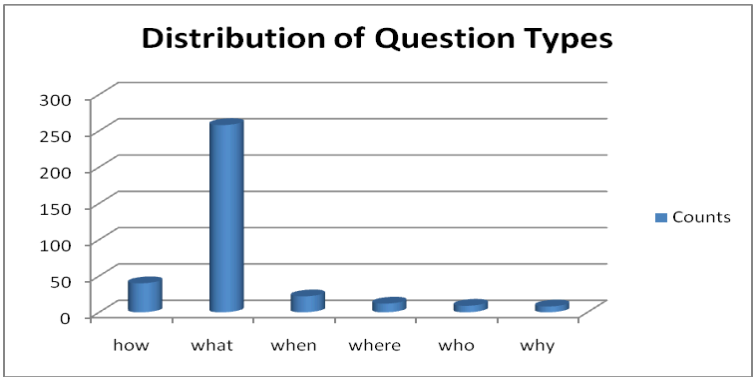

Figure 1. Distribution of Question Types.

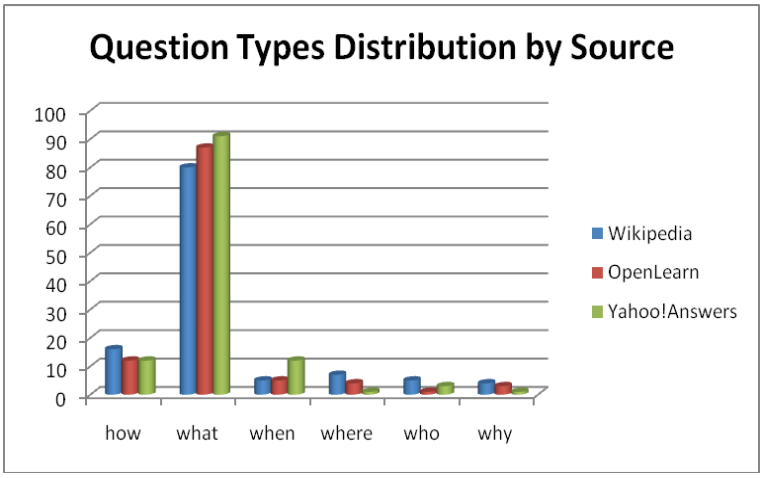

Figure 2. Distribution of Question Types by Input Paragraph Source.

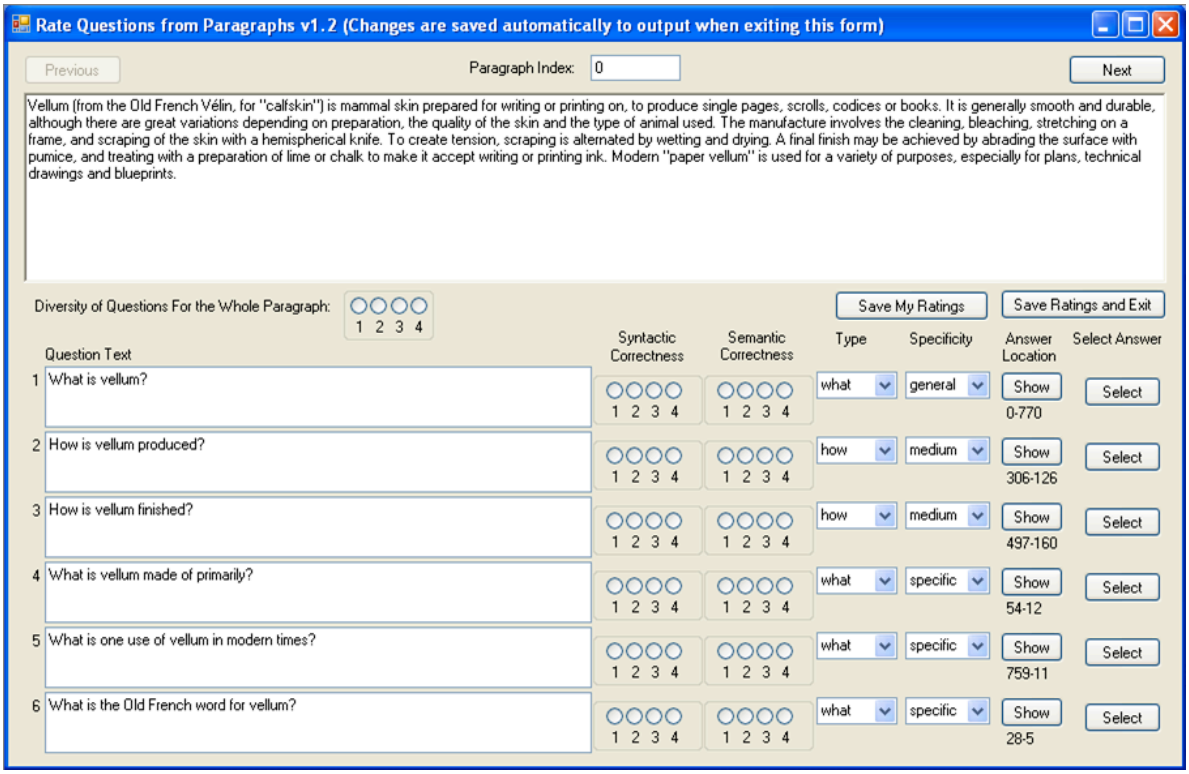

Figure 3. A screenshot of the Question Generation from Paragraphs rating tool.

3 Task B: Question Generation from Sentences

3.1 Task Definition

This task had four participants: Jadavpur University, University of Lethbridge, Saarland University, and University of Wolverhampton. Participants were given a set of inputs, with each input consisting of:

- a single sentence and
- a specific target question type (e.g., *Who?*, *Why?*, *How?*, *When?*; see below for the complete list of types used in the challenge).

For each input, the task was to generate 2 questions of the specified target question type. Input sentences, 90 in total, were selected from OpenLearn, Wikipedia and Yahoo! Answers (30 inputs from each source). Extremely short (<5 words) or long sentences (>35 words) were not included. Prior to receiving the actual test data, participants were provided with a development/example data set consisting of sentences from the aforementioned sources and, for one or more target question types, examples of questions. These questions were manually authored and cross-checked by the team organizing Task B. The three examples in Table 9 are taken from the development data set, one each from OpenLearn, Wikipedia and Yahoo! Answers.

Note that input sentences were provided as raw text. Annotations were not provided. There are a variety of NLP open source tools available to potential participants and the choice of tools and how these tools are used was considered a fundamental part of the challenge.

Table 9. Examples of sentences and corresponding questions from three sources.

Source	OpenLearn	
Sentence	The poet Rudyard Kipling lost his only son in the trenches in 1915.	
Target question type	*Who*	Who lost his only son in the trenches in 1915?
	When	When did Rudyard Kipling lose his son?
	How Many	How many sons did Rudyard Kipling have?
Source	Wikipedia	
Sentence	Two important variables used for the classification of igneous rocks are particle size, which largely depends upon the cooling history, and the mineral composition of the rock.	
Target question type	*Which*	Which two important variables are used for the classification of igneous rocks?
Source	Yahoo!Answers	
Sentence	In Australia you no longer can buy the ordinary incandescent globes, as you probably already know.	
Target question type	*Where*	Where can you no longer buy the ordinary incandescent globes?
	Yes/No	Can you buy the ordinary incandescent globes in Australia?
	What	What can you no longer buy in Australia?

Participants were also provided with the following list specifying the target question types:

- *Who*?: The answer to the generated question is a person (e.g. Abraham Lincoln) or group of people (e.g. the American people) named in the input sentence.
- *Where*?: The answer to the generated question is a placename (e.g. Dublin, Mars) or location (North-West, to the left of) which is contained in or can be derived from the input sentence.
- *When*?: The answer is a specific date (e.g. 3rd July 1973, 4th July), time (e.g. 2:35, 10 seconds ago), era or other representation of time.
- *Which*?: The answer will be a member of a category (e.g. Invertebrate or Vertebrate) or group (e.g. Colours, Race) or a choice of entities (e.g. Union or Confederacy) given in the input sentence.

- *What*?: The question might describe a specific entity mentioned in the input sentence and ask what it is. The question may also ask the purpose, attributes or relations of an entity as described in the input sentence.
- *Why*?: The question asks the reasoning behind some statement made in the input sentence
- *How many/long*?: The answer will be a duration of time or range of values (e.g. 2 days) or a specific count of entities (e.g. 32 counties) within the input sentence.
- *Yes/No*: The generated question should ask whether a fact contained in the input sentence is either true or false (e.g. Are mathematical co-ordinate grids used in graphs?).

3.2 Guidelines for Human Judges

The evaluation criteria fulfilled two roles. Firstly, they were provided to the participants as a specification of the kind of questions that their systems should aim to generate. Secondly, they also played the role of guidelines for the judges of system outputs in the evaluation exercise.

For this task, five criteria were identified:

- Relevance
- Question Type
- Syntactic Correctness and Fluency
- Ambiguity
- Variety

All criteria are associated with a scale from 1 to *N* (where *N* is 2, 3 or 4), with 1 being the best score and *N* the worst score. The criteria are defined as follows.

3.2.1 Relevance

Questions should be relevant to the input sentence. This criterion measures how well the question can be answered based on what the input sentence says.

Table 10. Scoring rubric for relevance.

Rank	*Description*
1	The question is completely relevant to the input sentence.
2	The question relates mostly to the input sentence.
3	The question is only slightly related to the input sentence.
4	The question is totally unrelated to the input sentence.

3.2.2 Question Type

Questions should be of the specified target question type.

Table 11. Scoring rubric for Question Type.

Rank	*Description*
1	The question is of the target question type.
2	The type of the generated question and the target question type are different.

3.2.3 Syntactic Correctness and Fluency

The syntactic correctness is rated to ensure systems can generate grammatical output. In addition, those questions which read fluently are ranked higher.

Table 12. Scoring rubric for syntactic Correctness and Fluency.

Rank	*Description*	*Example*
1	The question is grammatically correct and idiomatic/natural.	In which type of animals are phagocytes highly developed?
2	The question is grammatically correct but does not read as fluently as we would like.	In which type of animals are phagocytes, which are important throughout the animal kingdom, highly developed?
3	There are some grammatical errors in the question.	In which type of animals **is** phagocytes, which are important throughout the animal kingdom, highly developed?
4	The question is grammatically unacceptable.	**On** which type of animals **is** phagocytes, which are important throughout the animal kingdom, developed?

3.2.4 Ambiguity

The question should make sense when asked more or less out of the blue. Typically, an unambiguous question will have one very clear answer.

Table 13. Scoring rubric for Ambiguity.

Rank	*Description*	*Example*
1	The question is unambiguous.	Who was nominated in 1997 to the U.S. Court of Appeals for the Second Circuit?
2	The question could provide more information.	Who was nominated in 1997?
3	The question is clearly ambiguous when asked out of the blue.	Who was nominated?

3.2.5 Variety

Pairs of questions in answer to a single input (i.e., with the same target question type) are evaluated on how different they are from each other. This rewards those systems which are capable of generating a range of different questions for the same input.

Table 14. Scoring rubric for Variety.

Rank	*Description*	*Example*
1	The two questions are different in content.	Where was X born?, Where did X work?
2	Both ask the same question, but there are grammatical and/or lexical differences.	What is X for?, What purpose does X serve?
3	The two questions are identical.	

Each of the criteria is applied *independently* of the other criteria to each of the generated questions. Scoring on the criteria was done by two judges for each generated question (with the judges not knowing which system generated the question). For each question, one of the judges was a member of the QG-STEC team and the other a member of one of the participating teams

(though they never got to rate questions from their own system). In short, we had a mix of independent judges and peer-review. The scores of the two judges for each item were averaged, with the average score being used in the further calculations reported below.

3.3 Results & Discussion

This section contains a presentation and analysis of the results of the QG from sentences task of QG-STEC 2010.[9] The results concern a variety of criteria in order to determine possible improvements which might be made to both the QG from sentences task and to future QG systems.

The first of these criteria is the total number of questions generated for a data instance by all QG systems. The results for the rating criteria used by the human evaluators were also analysed both in relation to the different sources and systems. The marking system employed used lower scores for better questions. The best score for all criteria was 1. Thus, in the tables and charts used in this report the lower values indicate a better score.

3.3.1 Number of questions generated by Resource

The QG from sentences task used 30 sentences from each of three online resources, OpenLearn, Wikipedia and Yahoo! Answers, giving a total of 90 sentences. For each sentence one or more target question types was specified resulting in 180 possible generated questions. In order to measure the ability of a QG system to generate a variety of questions, participants were asked to generate two questions for each target sentence and type. The maximum number of submitted questions possible was then 360 times the number of participants (4) resulting in a total of 1440.

No. of Sentences		*Average no. of Question Types*		*Two questions per Question Type*		*Possible # of questions per participant*
90	**X**	2	**X**	2	=	360

Analysing the actual number of questions generated with regard to the online resource can provide some indication as to which resources are more difficult to work with from a QG perspective. Table 15 below shows the actual questions generated by participants for input sentences generated from each of the online resources.

Table 15. Percentage of questions generated per data resource.

Resource	Max achievable questions per resource	Actual generated questions per resource	Percentage generated
OpenLearn	480	312	65%
Wikipedia	480	309	64%
Yahoo!Answers	480	275	57%

All systems followed the guidelines returning up to two questions for each question type. The number of generated questions varied across the systems resulting in 62% of all requested questions. For sentences extracted from either the OpenLearn or Wikipedia resource it was found that systems generated approximately 65% of the maximum number of questions possible. The Yahoo! Answers resource appears to have been more challenging. For sentences extracted from this resource the systems only generated 57% of the maximum possible. The resource does contain many spelling errors and other language imperfections which make it difficult for NLP and this was also reflected in the STEC results.

[9] The raw data with the results can be found at http://computing.open.ac.uk/coda/data.html

3.3.2 Number of questions generated by Question Type

The target question types were specified for each sentence with the intention of discovering those types which provided more of a challenge for current QG systems. Analysing the results for actual questions generated per question type provides an indication of this. Table 16 below shows the percentage of actual generated questions per question type.

Table 16. Percentage of questions generated per question type.

	Max achievable questions per type	Actual generated questions per type	Percentage generated
When	144	107	74 %
Who	120	84	70 %
Where	120	74	62 %
What	464	285	61 %
Which	176	108	61 %
Why	120	72	60 %
How many	184	106	58 %
Yes/No	112	60	54 %

Assuming that the number of actually generated questions provides some indication of the ability to generate questions for a specific type, the data suggest that questions of some types are easier to generate than others. Participating systems generated 70% and above of the maximum possible sentences for 'when' and 'who' question types. For 'yes/no' question types it was found that only 54% of the maximum possible sentences were generated. This could indicate that 'yes/no' questions are more difficult to generate. This particular figure is, however, skewed because one of the participating systems had no rules at all for generating Yes-No questions and therefore never generated this type of question.

3.3.3 Average Scores by Resource

Questions were rated by our judges using five criteria. These criteria were then used to measure the performance of systems with regard to the criteria and to determine areas where QG systems might be improved. We first present and discuss the scores for the questions organised by resource.

Analysing the scores on the criteria with regard to the resource should again indicate any difficulty a particular resource might present, in this case according to a specific criterion. We also again analyse the results according to the target question type to discover any relation between target question types and the evaluation criteria.

Table 17 and Figure 4 below show the average scores for each criterion for each resource used in the STEC. Lower scores are better.

Table 17. Percentage of questions generated per question type.

	Relevance	**Question Type**	**Correctness**	**Ambiguity**	**Variety**
OpenLearn	1.54	1.07	2.14	1.54	1.65
Wikipedia	1.61	1.12	2.24	1.61	1.89
Yahoo! Answers	1.57	1.12	2.23	1.72	2.04

Figure 4. Scores for each question data source.

The data show that the sentences from the OpenLearn resource achieved as good as or better generation results than the other two resources on all criteria. One possible reason for this is that OpenLearn study units are rigorously checked before being published.

Question type, which ensures that generated questions are of the type requested by the test instance, is close to 1 (the optimal value) for all resources. In general, most generated questions were of the type specified.

Input sentences derived from Yahoo! Answers reflect the nature of that resource. Text from Yahoo! Answers contains a lot of vague, badly spelled, ambiguous and generally imperfect language.

3.3.4 Average Scores by Question Type

The evaluation criteria can also be used together with the various target question types in order to determine whether particular question types present difficulties for generation systems. Table 18 and Figure 5 below show the average scores for the five criteria according to the target question type for the data instance.

Table 18. Scores for each question type.

	Relevance	Question Type	Correctness	Ambiguity	Variety
How many	1.41	1.1	2.44	1.51	1.98
Who	1.45	1.06	1.88	1.48	1.57
When	1.54	1.08	2.37	1.72	1.87
Yes/No	1.57	1.05	2.35	1.63	2.02
What	1.58	1.08	1.98	1.63	1.68
Which	1.64	1.22	2.36	1.51	1.99
Where	1.66	1.05	2.36	1.69	1.99
Why	1.74	1.15	2.31	1.83	2.16

Figure 5. Scores for each question type.

There is variation in the results for the different criteria when grouped by target question type. The type 'how many' achieved the best results for matching question type and also the worst result for 'correctness'. 'Why' question types were amongst the poorest rated generated questions for all criteria. 'Who' question types achieved the best results in most criteria.

3.3.5 Average Scores by System

Five systems entered this task: MRSQG Saarland (Yao & Zhang, 2010), WLV Wolverhampton (Varga & Ha, 2010), JUGG Jadavpur (Pal et al., 2010) and Lethbridge (Ali et al., 2010). The averaged results for the systems on each of the evaluation criteria are depicted in Figure 6, with lower values indicating better scores. WLV scores best on all criteria except for "Variety". The picture changes when systems are penalized for missing questions (Figure 7), i.e., when a missing question is given the lowest possible score on each of the criteria. Now, MRSQG outperforms the other systems on all criteria.

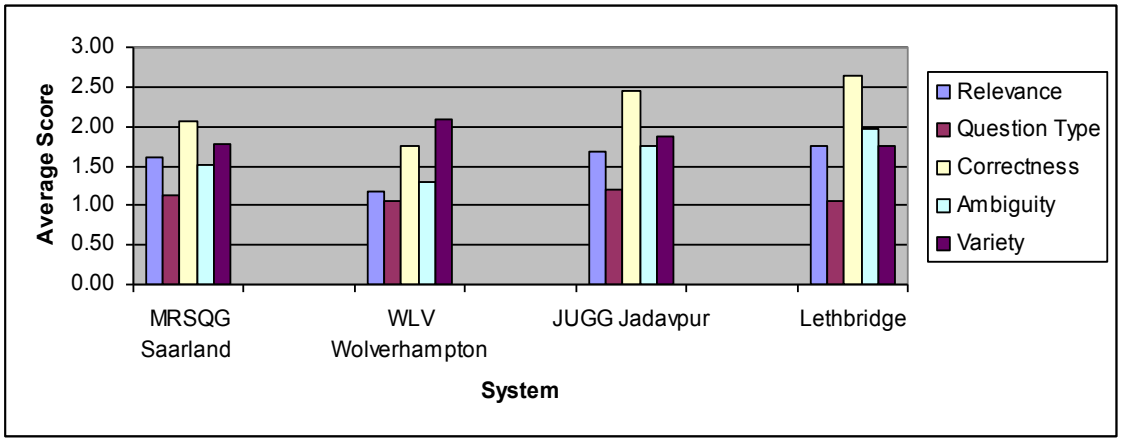

Figure 6. Results for QG from Sentences (without penalty for missing questions)

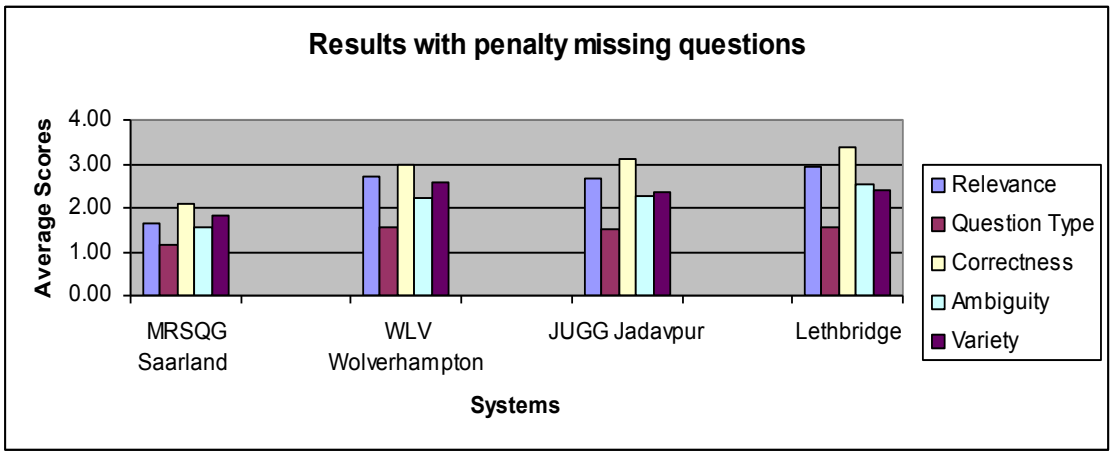

Figure 7. Results for QG from Sentence (with penalty for missing questions)

The Lethbridge system uses syntactic parsing, Part Of Speech (POS) tagging and a Named Entity analyzer to decorate the input sentence with annotations. Based on the syntactic parse, the sentence is simplified, dividing it into short sentences. These are then matched with rules. A rule consists essentially of a template that is matched against the input and corresponding question patterns. If the template is successfully matched with the structure of an input sentence, the result is one or more instantiated question patterns, i.e, questions. A similar approach is followed by JUQGG, though it uses a slightly different set of tools for decorating the input text with structure (e.g., it includes also semantic role labelling). There is also no sentence simplification step. There is, however, as with the Lethbridge system, a stage where the processed input is matched with templates and then mapped to question patterns - an approach also followed previously by, for example, Wyse & Piwek (2009). The WLV system also uses this approach. It is, however, set apart from the two other systems by not having any rules for Yes-No questions and including a checking stage where tests are performed on the input (does it contain negation, complex structures, etc.) which determine whether a question of sufficient quality can be generated. Finally, MRSQG forms a class of its own. It uses Natural Language Understanding technology to produce a language-independent deep semantic representation of the input sentence. The representation of the declarative content of the input sentence is transformed, using rules, to a representation of the content of a question. It then takes this question representation and uses an existing Natural Language Generation system to construct the surface realization of the question.

When we penalize systems for missing questions, the semantics-based deep approach used by MRSQG, on the QG-STEC data, outperforms the rule-based shallow approach of the other three systems. Interestingly, if we compute the scores using only the values for the actually generated questions, the WLV system comes out best. Here it seems that the WLV system strategy of checking the input prior to generation (and not generating an output if the input doesn't satisfy certain criteria) pays off.

4 Lessons Learned

The first QG-STEC was definitely a success by many measures including number of participants, results, and resources created, given the lack of funded projects on the topic. As already mentioned, there was a big discrepancy between number of researchers who expressed their interest to participate and the number of researchers who actually participated. In the absence of research funding targeting QG research efforts, researchers could not afford spending substantial effort on developing competitive algorithms to participate in QG-STECs. Funding of systematic research on QG is desperately needed to advance the field and keep the momentum going.

After reflecting on the process and outcomes, we have identified several aspects that have helped making the task a success and some that could be improved in the next round of shared evaluations of tasks A and B or that could be used in planning other QG tasks.

One key aspect of process to put a QG-STEC together was the involvement of the QG community. The community was asked to submit candidate QG tasks for a first QG-STEC. Then, the community at large was polled regarding which tasks to choose for the first QG-STEC. We believed that more than two tasks would have diluted the success of a first QG-STEC. By polling the community, we indirectly assured maximum participation. One downside of the community polling process was the introduction of some delays regarding the official announcement of the QG tasks to be offered in the first QG-STEC. We recommend the next QG-STEC should be announced at least one year in advance of the expected publication of the results. This will give enough lead time to advertise the event and also give participants sufficient time to get ready for the challenge. In terms of planning, we recommend that the test period be scheduled at least 4 months before the publication of the results such that a peer-evaluation process can be effectively implemented.

One other positive aspect of the first QG-STEC was the post-challenge group discussions at the 3rd Workshop on Question Generation where the results of the STEC were made publicly available. The participants had a chance to provide suggestions for improving the next round of the STEC. For instance, the participants recommended that for Task A, Question Generation from Paragraphs, we eliminate Yahoo! Answers as a source of texts due to language challenges posed by unedited Yahoo! Answers paragraphs. Indeed, we compared means of the five quality criteria with source of input paragraphs as the factor (N=117 for Wikipedia; N=112 for OpenLearn; N=120 for Yahoo! Answers). A significant difference was found for the syntactic complexity criterion ($p=0.015<0.05$). A post-hoc Tukey test revealed significant differences between Wikipedia and Yahoo! Answers groups ($p=0.013<0.05$). As replacement for Yahoo! Answers, there was a suggestion to consider texts from domains that are of high-interest, for example, in biomedical domains. This may attract research groups working on biomedical texts, which could give a boost to the QG-STEC and QG research efforts in general. Another suggestion was to not only ask participants to generate questions and indicate the fragment in the input paragraph that triggered the question, but also to ask participants to generate the answers themselves. This latter suggestion needs more thought before it is implemented, as it seems to add another level of complexity to the QG-STEC.

One last general suggestion is to consider replacing rating scales, which we used in evaluating the submissions of the first QG-STEC, with preference judgments as the former seem to pose some challenges. They may be unintuitive for raters and the inter-rater agreement tends to

be low when using rating scales (Belz & Kow, 2010a). A rating scale makes raters judge questions in absolute terms. In contrast, preference judgments would allow them to judge which of two items they prefer for a particular evaluation criterion, e.g. syntactic correctness. If preference judgments are implemented in a future QG-STEC, raters will be shown several questions at a time, submitted by different participants, and asked to rank them based on each of the criteria. Belz and Kow (2010b) showed that using preference judgments for language generation tasks leads to better inter-annotator agreement and also explains a larger proportion of variation accounted for by system differences.

We also have some specific recommendations for running the next round Tasks of A and B. A spin-off of Task A could be a task focusing on content selection as this step seems to be the more challenging for this task as it involves discourse level processing. A user model can be added to such a task in order to make it more interesting. Furthermore, a task on content selection can target a specific application such that there is a clear set of goals that could be used to quantify the importance of various content items. Other spin-off tasks can focus on question type selection and question construction as well.

For Task A, there was on approach proposed in which many questions are first generated from a myriad of potential content items in the input paragraph followed by a ranking phase. The importance of a question was determined in two separate steps. First, predicate argument structures along with semantic roles are used to identify important aspects of paragraphs. This step has limitations when it comes to generating questions that should rely on cross-sentence information, e.g. a cause-effect discourse relation between two sentences. Second, the approach preferred questions generated from main clauses and questions that do not include pronouns because no coreference resolution was used. The question formulation step used the selected content and corresponding verb complex in a sentence together with a set of reformulation rules to generate the actual questions. They avoided using discourse parsers in their approach due to the modest current performance of these parsers. We will have to provide discourse annotations for the input paragraphs in a future offering of Task A, either by manually improving the output of available discourse parsers, such as HILDA, or by manually annotating the input paragraphs. While this may seem inconsistent with our principle of using raw input for the task, we feel this minimal annotation of input for discourse relations is necessary given the young stage of development of discourse parsers relative to the maturity of other standard language processing steps such as part of speech tagging or syntactic parsing.

For Task B, there were essentially two approaches: a ‘shallow’ and a ‘deep’ one. Whereas the shallow approach relied on a combination of syntactic and semantic annotations of the input to guide rules for mapping declarative text to questions, the deep approach relied on full computation of a language-independent semantic representation of the input which is mapped to a semantic representation of the output, which is then mapped in turn to a question using language generation technology. The deep approach (as used by the MRSQG system) performed best, provided that missing questions were penalized (i.e., questions that the system was asked to generate but did not generate). One of the ‘shallow’ systems, WLV, did however come out best when missing questions were not taken into account. This particular system distinguished itself from the other systems by including a checking stage, where, based on an analysis of the input, the system decided whether to generate an output. We should also add that the WLV system could have performed even better if it had included rules for Yes-No questions (these were completely missing from this system).

There are several other ways to improve future runs of the two tasks offered in the first QG-STEC. A suggestion is to increase the number of input data items to at least 60 items per sources such that statistical significance tests could be performed to compare results across the various sources of input text.

Another suggestion is to add a Specificity field to the submission requirements. The Specificity field will require participants to indicate what the scope is for each of the questions

(general, medium, or specific). In the first QG-STEC, the scope was somehow inferred from the answer location range by the rater.

We also need to develop a smart scheme to combine the individual scores into an overall score for Tasks A and B. We believe a second run of the tasks may give us enough evidence to develop an appropriate overall score by weighting the various criteria according to some principles.

5 Conclusions

The first QG-STEC was a success in terms of participants and resources created (detailed results are not discussed due to space constraints, but see the links to the raw results data in the sections above for Task A and B). We now have several datasets of text-question pairs together with human judgments using various criteria. Also, we developed a software tool that can be used by humans to rate questions (see Figure 1). Participants have proposed and implemented several approaches that can be used as a starting point for future QG-STECs and sources of inspiration by newcomers. While these resources are valuable, there is still need for larger datasets, linguistic annotations for the input data, and more tools and improved evaluation methodologies. The short term goal should be scaling up the QG-STEC for tasks A (Question Generation from Paragraphs) and B (Question Generation from Sentences) such that more data is available to enable more solid results and more participants. In the long term, the QG-STEC should scale up in terms of tasks, e.g. more targeted tasks that focus on specific aspects of the Question Generation process such as question type selection. The survival and future successes of the QG-STEC movement and of the QG research community at large is highly dependent on the availability of funding resources directed specifically at Question Generation research projects.

Acknowledgments

We are grateful to a number of people who contributed to the success of the First Shared Task Evaluation Challenge on Question Generation: Rodney Nielsen, Amanda Stent, Arthur Graesser, Jose Otero, and James Lester. Also, we would like to thank the National Science Foundation who partially supported this work through grants RI-0836259 and RI-0938239 (awarded to Vasile Rus), the Institute for Education Sciences who partially funded this work through grant R305A100875 (awarded to Vasile Rus), and the Engineering and Physical Sciences Research Council who partially supported the effort on Task B through grant EP/G020981/1 (awarded to Paul Piwek). The views expressed in this paper are solely the authors'.

References

Ali, H, Chali, Y., Hasan, S. (2010). Automation of Question Generation from Sentences, In: Boyer & Piwek (2010), pp. 58-67.

Beck, J.E., Mostow, J., and Bey, J. (2004). Can Automated Questions Scaffold Children's Reading Comprehension? *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 478-490. 2004. Maceio, Brazil.

Belz, A., Kow, E., Viethen, J. and Gatt, A. (2008) The GREC Challenge 2008: Overview and Evaluation Results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG'08)*, pp. 183-191.

Belz, A. and Kow, E. (2010a). The GREC Challenges 2010: Overview and Evaluation Results. In *Proceedings of the 6th International Natural Language Generation Conference (INLG'10)*, pp. 219-229.

Belz, A. and Kow, E. (2010b). Comparing Rating Scales and Preference Judgements in Language Evaluation. In *Proceedings of the 6th International Natural Language Generation Conference (INLG'10)*, pp. 7-15.

Boyer, K. and Piwek, P. (2010) (Eds). QG2010: The Third Workshop on Question Generation, Carnegie Mellon University, Pittsburgh, PA, USA.

Dale, R. & M. White (2007) (Eds.). *Position Papers of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

duVerle, D. and Prendinger, H. (2009). A Novel Discourse Parser Based on Support Vector Machines. Proc 47th Annual Meeting of the Association for Computational Linguistics and the 4th Int'l Joint Conf on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09), Singapore, Aug 2009 (ACL and AFNLP), pp 665-673.

Edmonds, P. (2002). Introduction to Senseval. ELRA Newsletter, October 2002.

Judita Preiss and David Yarowsky (2001). Editors. *The Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems.*

Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J., and Oberlander, J. (2010). Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2). In Proceedings of the 6th International Natural Language Generation Conference (INLG), Dublin, Ireland.

Kunichika, H., Katayama, T., Hirashima, T., & Takeuchi, A. (2001). Automated Question Generation Methods for Intelligent English Learning Systems and its Evaluation, Proc. of ICCE01.

Lauer, T., Peacock, E., & Graesser, A. C. (1992) (Eds.). *Questions and information systems*. Hillsdale, NJ: Erlbaum.

Lin, C.Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, post-conference workshop of ACL 2004, Barcelona, Spain.

Lin, C. and Och, F. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics, In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*.

Mannem, P., Prasad, R., and Joshi, A. (2010). Question Generation from Paragraphs at UPenn: QGSTEC System Description, Proceedings of the Third Workshop on Question Generation (QG 2010), Pittsburgh, PA, June 2010.

Mitkov, R., Ha, L. A. and Karamanis, N. (2006): A Computer-Aided Environment for Generating Multiple-Choice Test Items. *Natural Language Engineering 12*(2). 177-194. Cambridge University Press.

Mostow, J. and Chen, W. (2009). *Generating Instruction Automatically for the Reading Strategy of Self-Questioning.* Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED2009), Brighton, UK, 465-472.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B., and Valeri, J. (2004). Using Automated Questions to Assess Reading Comprehension, Vocabulary, and Effects of Tutorial Interventions. *Technology, Instruction, Cognition and Learning*, 2004. *2*: p. 97-134.

Pal, S., Mondal, T., Pakray, P., Das, D., Bandyopadhyay, S. (2010). QGSTEC System Description: JUQGG: A Rule Based Approach. In: Boyer & Piwek (2010), pp.76-79.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation, in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318.

Piwek, P., H. Hernault, H. Prendinger, M. Ishizuka (2007). T2D: Generating Dialogues between Virtual Agents Automatically from Text. In: Intelligent Virtual Agents: Proceedings of IVA07, LNAI 4722, September 17-19, 2007, Paris, France, (Springer-Verlag, Berlin Heidelberg) pp.161-174.

Piwek, P. and S. Stoyanchev (2011). Data-Oriented Monologue-to-Dialogue Generation. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 242-247, Portland, Oregon, June 19-24, 2011.

Reiter, E. & Dale, R. (1997). Building Applied Natural-Language Generation Systems. *Journal of Natural-Language Engineering*, **3**:57-87.

Reiter, E. & Dale, R. (2000). *Building Applied Natural-Language Generation Systems,* Oxford University Press, 2000.

Rus, V., Cai, Z., Graesser, A.C. (2007a). Experiments on Generating Questions About Facts. Alexander F. Gelbukh (Ed.): Computational Linguistics and Intelligent Text Processing, 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007

Rus, V., Graesser, A.C., Stent, A., Walker, M., and White, M. (2007b). Text-to-Text Generation, in Shared Tasks and Comparative Evaluation in Natural Language Generation by Robert Dale and Michael White, November, 2007, pages 33-46.

Rus, V. and Graesser, A.C. (2009). *Workshop Report: The Question Generation Task and Evaluation Challenge*, Institute for Intelligent Systems, Memphis, TN, ISBN: 978-0-615-27428-7.

Varga, A. and Ha, L.A. (2010). WLV: A Question Generation System for the QGSTEC 2010 Task B. In: Boyer & Piwek (2010), pp. 80-83.

Voorhees, E. M. and Tice, D.M. (2000). The TREC-8 Question Answering Track Evaluation. In E.M. Voorhees and D.K. Harman, editors, Proceedings of the Eighth Text REtrieval Conference (TREC-8), pages 83-105, 2000. NIST Special Publication 500-246.

Walker, M., Rambow, O., & Rogati, M. (2002). Training a Sentence Planner for Spoken Dialogue Using Boosting, *Computer Speech and Language Special Issue on Spoken Language Generation* , July 2002.

Wolfe, J.H. (1976). "Automatic question generation from text - an aid to independent study." SIGCUE Outlook **10**(SI): 104--112.

Wyse, B. and P. Piwek (2009). Generating Questions from OpenLearn Study Units. In: V. Rus and J. Lester (Eds.), Proceedings of the 2nd Workshop on Question Generation, AIED 2009 Workshop Proceedings, pp. 66-73.

Yao, X and Zhang, Y. (2010). Question Generation with Minimal Recursion Semantics. In: Boyer & Piwek (2010), pp.68-75.